MTA Számitástechnikai és Automatizálási Kutató Intézete

Veleszületett rendellenességek közti kapcsolat megállapitása mátrixok szinguláris felbontásával

Bolla Marianna

Bevezetés

Jelen előadás az NJSZT 1978-as szegedi kollokviumán Bolla M., Czeizel E., Telegdi L., Tusnády G.: "Többszörös veleszületett rendellenességek statisztikai vizsgálata" cimmel elhangzott előadásához kapcsolódik. A feladat ujszülöttek veleszületett rendellenességei közti kapcsolat megállapitása volt. A minta a Magyarországon 1970-76. között született $N = 1\,186\,742$ ujszülöttből állt és $n = 40$ féle rendellenességet figyeltek meg rajtuk.

Egy olyan, a többdimenziós normális eloszláson alapuló küszöbmodellt ismertettünk, amelynek lényege, hogy a rendellenességek egyszeres $(0_T(i))$ és páros $(0_T(i,j))$ előfordulásai egyértelmüen meghatározzák a többszörös eseteket. Ui. jelölje A_i az i-edik rendellenességet és L_i a hozzá tartozó háttérváltozót /hajlamot/; tegyük fel, hogy ezek standard normális eloszlásu valószinüségi változók /vsz.v./ R kovarianciamátrixszal. A küszöbmodell szerint egy $G \subset \{A_1, A_2, \ldots, A_n\}$ rendellenességkombináció bekövetkezésének valószinüsége:

$$P(G) = P(L_k \geq T_k, A_k \in G) \qquad [1]$$

ahol T_k küszöbök az egyes rendellenességek populációs

gyakoriságából határozhatók meg. A rendellenességek $\Sigma = (\rho\ i\ j)_{i=1,j=1}^{n\ \ n}$, empirikus korrelációs mátrixát az

$$N \cdot P\ (L_i \geq T_i,\ L_j \geq T_j) = O_T\ (i,j), \quad 1 \leq i < j \leq n \qquad [2]$$

összefüggésekből becsültük.

A többdimenziós normális eloszlásra vonatkozó sorfejtéssel azt találtuk, hogy a modell egészen jó illeszkedést mutat a kettőnél több rendellenességet tartalmazó kombinációk esetén. Igy a rendellenességek strukturája a Σ empirikus korrelációs mátrix strukturájának elemzésén alapszik. Ennek legkézenfekvőbb módszere a faktoranalizis.

Faktoranalizis

L_i-ket közeliteni akarjuk $k < n$ számu független, standard normális eloszlásu valószinüségi változók lineáris kombinációjával. Jelölje ezeket $\underline{f} = (f_1,\ldots,f_k)$ és legyen $\underline{L} = (L_1,\ldots,L_n)$! A modell:

$$\underline{L} = \Lambda \underline{f} + \underline{e} \qquad [3]$$

ahol Λ $n \times k$-as mátrix, $\underline{e} = (e_1,\ldots,e_n)$, független komponensü, normális eloszlásu vsz.v.v., $\underline{f}$ a faktorokat, $\underline{e}$ pedig a változók egyedi részét és a mérési hibákat tartalmazza. Mivel e_i-k függetlenek, a rendellenességek közti kapcsolatok faktortérbeli vetületeikkel interpretálhatók, melyet Λ oszlopai feszitenek ki, mint bázisvektorok. A feladat mátrixelméleti ekvivalense Σ-nak.

$$\Sigma = \Lambda\Lambda' + \psi \qquad [4]$$

alaku felbontása, ahol Λ $n \times k$-as k-adrangu mátrix, ψ

pedig $n \times n$-es diagonális mátrix, diagonálisában pozitiv elemekkel.

A [4] mátrixegyenletet maximum likelihood módszerrel oldottuk meg, ahol a likelihood függvény logaritmusát

$$-\frac{N}{2}[\log|R| + \mathrm{tr}(\Sigma R^{-1})] - \frac{Nn}{2}\log(2\pi) \qquad [5]$$

maximalizáltuk azzal a mellékfeltétellel, hogy $\Lambda\psi^{-1}\Lambda$ diagonális mátrix.

Rögzitett ψ mellett Λ-ban, majd a kapott Λ rögzitése esetén ψ-ben minimalizáltunk. Az eljárást többször megismételve egy *iterativ* algoritmust kapunk, ahol kezdőértékeknek $\psi^{(o)}$ diagonálisában az $1-\frac{1}{\sigma^{ii}}$ értékeket választottuk, ahol $(\sigma^{ij})_{i=1,j=1}^{n\quad n} = \Sigma^{-1}$ /ami éppen az i-edik változónak a többivel vett többszörös korrelációs együtthatója. A minimalizálás során a parciális determináltakat 0-vá téve, $\psi^{(i)}$ $(i = 0,1,\ldots,M)$ rögzitése esetén $\Lambda^{(i+1)}$-re a következő megoldást kapjuk:

$$\Lambda^{(i+1)} = \psi^{(i)^{1/2}}\ \Omega(\Theta - I)^{1/2} \qquad [6]$$

ahol Θ tartalmazza $\psi^{(i)^{-1/2}}\ \Sigma\psi^{(i)^{-1/2}}$ szinguláris értékeit nagyság szerint csökkenő sorrendben, Ω pedig a hozzájuk tartozó normált s.v.-okat.

$$\psi^{(j+1)} = \mathrm{diag}(\Sigma - \Lambda^{(j)}\Lambda^{(j)'}), \quad j = 1,\ldots,M-1 \qquad [7]$$

ahol $\mathrm{diag}(A)$ az A mátrix fődiagonálisát tartalmazó diagonális mátrix.

Az iterációs lépések során k értéke természetesen változik /esetünkben csökkent/. [6]-ban egy mátrix szingu-

láris felbontását végezzük el, az egyes főkomponenseket egymás után leválasztva. Ott a megállási kritérium az, hogy az egyes változók kommunalitásai meghaladjanak egy bizonyos $1-\varepsilon$ szintet, ahol $0 < \varepsilon < 1$ előre adott konstans. Ezenkivül minden egyes iterációban vizsgáljuk a

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{[O_T(i,j)-N \cdot p_k(i,j)]^2}{N \cdot p_k(i,j)} \qquad [8]$$

likelihood alakulását is, ahol $p_k(i,j)$ A_i és A_j együttes előfordulásának valószinüsége a k faktoros modellben, és az A_i, A_j közötti korreláció faktormodellbeli értékéből számolható, ami

$$\sum_{\ell=1}^{k} \lambda_{i\ell} \lambda_{j\ell} \qquad (1 \leq i < j \leq n). \quad [9]$$

Az iterációs lépések M számát idő- és hibakorlátok mellett még az a - konkrét feladatból vett - mellékfeltétel is meghatározza, hogy a [8] likelihood értéke kisebb legyen, mint az

$$\frac{n(n+1)}{2} - \left[nk - \frac{k(k-1)}{2}\right] \qquad [10]$$

szabadsági foku X^2 0,95-ös szignifikanciaszinten.

Mátrixok szinguláris felbontása

Az előbbi iteráció minden egyes lépésében szükség volt egy szimmetrikus mátrix szinguláris felbontására. Tegyük fel, hogy a mátrixnak nincsenek 0 sajátértékei, hiszen akkor az eredmények már egy alacsonyabb dimenziós térben is interpretálhatók. Ha mátrixunk nem pozitiv definit, a szinguláris értékek nem egyeznek meg a sajátér-

tékekkel, hanem azok abszolut értékével. A szimmetricitás miatt azonban a bal- és jobboldali sajátvektorok esetleges (-1)-es szorzótól eltekintve egyenlőek.

Tegyük fel továbbá azt is, hogy a szinguláris értékek különbözőek; ugyanis egyenlő szinguláris értékek esetén a sajátvektorok szabadon választhatók egy - a multiplicitással egyenlő dimenziójú - altéren belül.

Legyen tehát A valós, szimmetrikus mátrix $\delta_1 > \ldots > \delta_n > 0$ szinguláris értékekkel. Akkor A /a sajátvektorok irányát rögzitve/ egyértelmüen előáll

$$A = U' \Delta V \qquad [11]$$

alakban, ahol Δ jelenti a szinguláris értékeket -, U a hozzájuk tartozó sajátvektorokat, V pedig ezek ± 1-szeresét oszloponként tartalmazó mátrix.

A hatványiteráció-, illetve inverz hatványiteráció módszerénél, /ami a sajátértékek és sajátvektorok direkt meghatározását célozza/, esetünkben sokkal gyorsabbnak bizonyultak azok a módszerek, amelyek az eredeti mátrixot egyszerübb alakra hozzák hasonlósági transzformációval.

Szimmetrikus mátrixok esetén a kontinuáns alakra hozás sokkal gyorsabbnak bizonyult a Hessenberg-féle alakra hozásnál. Az előbbi mátrix sajátértékeit az un. Gersgorin-körökön belül lokalizáltuk, majd a Sturm-tétel segitségével, intervallumfelezéssel teljesen behatároltuk őket. A második esetben a kapott Hessenberg-féle mátrixon QR--transzformációt hajtottunk végre. Megfelelő eltolások alkalmazásával az eljárás gyorsan konvergált.

A korrelációs mátrix szinguláris értékei

1.	10.8831	21.	0.7658
2.	2.9414	22.	0.6735
3.	2.2750	23.	0.6542
4.	1.9341	24.	0.6469
5.	1.8732	25.	0.6069
6.	1.6608	26.	0.5659
7.	1.6184	27.	0.4978
8.	1.5719	28.	0.4018
9.	1.4412	29.	0.3720
10.	1.3171	30.	0.3556
11.	1.2832	31.	0.2946
12.	1.1964	32.	0.2744
13.	1.1697	33.	0.2568
14.	1.0732	34.	0.2424
15.	1.0328	35.	0.1799
16.	0.9767	36.	0.1360
17.	0.8813	37.	0.0911
18.	0.8799	38.	0.0895
19	0.8478	39.	0.0243
20.	0.8235	40.	0.0233

I. táblázat

Programfuttatási tapasztalatok

A programokat az MTA SZTAKI CDC 3300-as számitógépén futtattuk. Az összehasonlitásban problémát okozhat, hogy az első módszerre FORTRAN, a másodikra pedig SIMULA programot alkalmaztunk. Az első módszer időigénye 39, utóbbié 71 sec volt. A szinguláris értékek 10^{-4}, a sajátvektorok pedig 10^{-3} pontossággal megegyeztek. /Lásd I. táblázat./

A konkrét feladatra konstruált likelihood alapján 13 szignifikáns faktort találtunk. Az első faktorhoz tartozó faktor-szkórok nagyságrendileg megfelelnek az egyes rendellenességek gyakoriságának a beteg populációban. Igy az első faktort lehetne az "immungyengeség" faktorának nevezni, a többi faktornak nem tudtunk ilyen szemléletes jelentést tulajdonitani.

Általában is elég nehéz előzetes hipotézisek nélkül végrehajtott faktoranalizis eredményeit interpretálni. Eddig nálunk a faktoranalizis inkább dimenziócsökkentési és strukturaegyszerüsitő eljárásként szolgált, ami lehetővé teszi nagyszámu függő változó kevesebb független változóval való helyettesitését és igy többdimenziós valószinüségek gyors, tömeges kiszámolását. Ha lesznek előzetes feltételezéseink a faktorok jelentésére vonatkozóan, próbálkozni fogunk szelektiv faktoranalizissel is.

Irodalomjegyzék

(1) Bolla M., Czeizel E., Telegdi L., Tusnády G.: Többszörös veleszületett rendellenességek statisztikai vizsgálata. NJSZT 9. Kollokvium, Szeged, 1978.

(2) Noble: Applied linear algebra, Prentice Hall, Inc. Englewood Cliffs, New Jersey.

(3) Anderson, T.W.: An introduction to multivariate statistical analysis. New York, 1958, John Wiley and Sons.

(4) Tusnády G., Csiszár A., Telegdi L., Czeizel E., Bolla M.: Statistical study of the multiple congenital malformations in Hungary, 1970--76, Transactions of the VIII Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Volume 8, 301-308.